Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation

Year: 2023 | Citations: 423 | Authors: Natalia Díaz Rodríguez, J. Ser, Mark Coeckelbergh, Marcos L'opez de Prado, E. Herrera-

Abstract

Trustworthy Artificial Intelligence (AI) is based on seven technical requirements sustained over three main pillars that should be met throughout the system's entire life cycle: it should be (1) lawful, (2) ethical, and (3) robust, both from a technical and a social perspective. However, attaining truly trustworthy AI concerns a wider vision that comprises the trustworthiness of all processes and actors that are part of the system's life cycle, and considers previous aspects from different lenses. A more holistic vision contemplates four essential axes: the global principles for ethical use and development of AI-based systems, a philosophical take on AI ethics, a risk-based approach to AI regulation, and the mentioned pillars and requirements. The seven requirements (human agency and oversight; robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability) are analyzed from a triple perspective: What each requirement for trustworthy AI is, Why it is needed, and How each requirement can be implemented in practice. On the other hand, a practical approach to implement trustworthy AI systems allows defining the concept of responsibility of AI-based systems facing the law, through a given auditing process. Therefore, a responsible AI system is the resulting notion we introduce in this work, and a concept of utmost necessity that can be realized through auditing processes, subject to the challenges posed by the use of regulatory sandboxes. Our multidisciplinary vision of trustworthy AI culminates in a debate on the diverging views published lately about the future of AI. Our reflections in this matter conclude that regulation is a key for reaching a consensus among these views, and that trustworthy and responsible AI systems will be crucial for the present and future of our society.